<Bioinformatics Center>Chemical Life Science

Bioinformatics Center
– Chemical Life Science –

http://cls.kuicr.kyoto-u.ac.jp

Prof
OGATA, Hiroyuki
(D Sc)

Assoc Prof
GOTO, Susumu
(D Eng)

Assist Prof
BLANC-MATHIEU, Romain
(D Sc)

Program-Specific Res
YOSHIZAWA, Akiyasu
(D Sc)

Researchers(pt)

SHIMIZU, Yugo
YAMAMOTO, Rumiko

Lect (pt)

MOGUSHI, Kaoru (D Sc) Juntendo University

Students

NISHIMURA, Yosuke (D3)
MIHARA, Tomoko (D3)
YOSHIKAWA, Genki (M1)
NISHIYAMA, Hiroki (M1)
KUMABE, Akihiko (UG)
ABDUL HALIM, Arhaime Hamezze (UG)
KHAIRIL ANUAR, Mohammad Fahmi Arieef (UG)
MOHD HASRI, Nurul Nadzirah (UG)
MUHAMAD, Rafidah (UG)

Scope of Research

We are interested in understanding the functioning and evolution of biological systems at varying scales from tiny microbes up to the Earth's environment, by leveraging rapidly accumulating big data in life science and bioinformatics approaches. We currently focus on 1) the evolution of viruses and their links to the origin of life, 2) microbial ecology in different ecosystems, and 3) the development of bioinformatics methods and biological knowledge resources for biomedical and industrial applications. To fuel these research activities, we take part in environmental sampling campaigns such as *Tara Oceans*. Our resources and developed tools are accessible through GenomeNet (www.genome.jp) to scientific communities and the public.

KEYWORDS

GenomeNet
(Meta)genomics
Pharmacoinformatics
Bioinformatics
Evolutionary Biology

Selected Publications

Ogata, H.; Takemura, M., A Decade of Giant Virus Genomics: Surprising Discoveries Opening New Questions, *In "Global Virology I - Identifying and Investigating Viral Diseases", Shapshak, P. et al., Eds., Springer, New York Heidelberg Dordrecht London*, 147-160 (2015).

Takemura, M.; Yokobori, S.; Ogata, H., Evolution of Eukaryotic DNA Polymerases via Interaction between Cells and Large DNA Viruses, *J. Mol. Evol.*, **81,** 24-33 (2015).

Sunagawa, S., *et al.*, Structure and Function of the Global Ocean Microbiome, *Science*, **348,** 1261359 (2015).

Brum, J. R., *et al.*, Patterns and Ecological Drivers of Ocean Viral Communities, *Science*, **348,** 1261498 (2015).

Mihara, T.; Goto, S.; Ogata, H., Diversity and Ecology of Marine Giant Viruses Uncovered from Their Genomes, *Seibutsu no Kagaku, Iden*, **69,** 318-325 (2015) (in Japanese).

Phylogenetic Diversity of Multi-domain Type III Polyketide Synthases in Protists

Polyketides (PKs) are natural products that have diverse chemical structures and biological functions as well as a wide variety of pharmacologically valuable properties. PKs are biosynthesized by polyketide synthases (PKSs) which are classified into three types (I, II, and III) according to their domain structures and subunit organizations. Type III PKSs consist of only a single ketosynthase (KS) domain whereas type I and II PKSs are either multi-domain proteins or protein complexes. Exceptional multi-domain type III PKSs, which contain type I PKS domains and catalyze both type I and III PKS reactions consecutively, have been isolated from a social amoeba (*Dictyostelium discoideum*).

We extracted 1,044 type III PKSs by homology-based genome mining against KEGG GENES to investigate the diversity of the type III PKS. While 1,034 type III PKSs from bacteria, plants, and fungi are single domain proteins, only seven out of ten type III PKSs from the other mostly unicellular eukaryotes (called "protists") are multi-domain proteins. Since the sequenced protist genomes are still taxonomically biased and limited, we further surveyed sequences of the Marine Microbial Eukaryote Transcriptome Sequencing Project, which represents a large number of protist transcriptome sequences. A phylogenetic tree of the resulting 307 protist sequences showed five distantly related clades, suggesting that the protist type III PKSs are more diverse than previously recognized (Figure 1). Among the five clades, four clades contain multi-domain proteins with novel domain organization. Therefore, multi-domain type III PKSs are frequent in and unique to protists. The function of these novel PKSs should be elucidated.

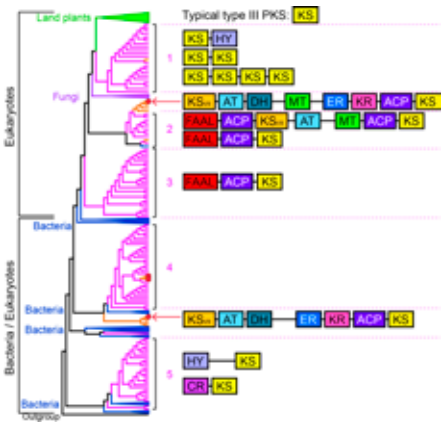

Figure 1. Phylogenetic tree and domain organization diversity of protist type III PKSs. The maximum likelihood tree based on KS domain sequences in type III PKSs and typical multi-domain structures for each clade are indicated. Sequences from genomic data and transcriptomic data are colored orange and magenta, respectively. Experimentally characterized sequences are indicated by red circles.

Investigation of Binding Sites of Arabidopsis Response Regulator Through Chromatin Immunoprecipitation Sequencing

Arabidopsis response regulator 1 (ARR1) is a response regulator in His-Asp phosphorelay. The motif 5'-GAT(C/T)-3' (referred to as the "core motif") is essential for ARR1 to bind to DNA strands. Furthermore, the extended version of the motif, 5'-AAGAT(C/T)TT-3' ("extended core motif"), appears more frequently in promoters directly regulated by ARR1 than in randomly chosen promoters. However, the genomic context of ARR1-binding sites is yet to be fully elucidated.

In this perspective, we used chromatin immunoprecipitation sequencing (ChIP-Seq) data of non-treated (control), water-treated (negative control), and benzyl adenine (activator of ARR1) –treated *Arabidopsis thaliana* (*At*) plants to investigate the binding sites of ARR1. These datasets and *At* reference genome data were used for normalization, and to determine a set of DNA sequences that are preferentially bound by ARR1.

Our initial analysis by MEME performed on the raw dataset did not yield any strongly conserved motifs due to complexity of the data. Therefore, we attempted to reduce data complexity by analyzing the relationship between different variables. As a result, a correlation between sequence length and quality score was found (Figure 2A). We also observed that sequences with scores above 40 contain the reference core motif at a relatively high density (Figure 2B). This result encourages us to continue to refine our dataset to achieve reliable detection of ARR1 binding site features. Alternatively, we are thinking about applying the Motif Centrality Analysis of ChIP-Seq (MOCCS) tool to clarify DNA-binding motif ambiguity. This method comprehensively analyzes and describes the frequency of every k-mer around the binding sites (bound by ARR1 in our case) determined as "peaks" by mapping the ChIP-Seq reads onto the *At* reference genome sequence.

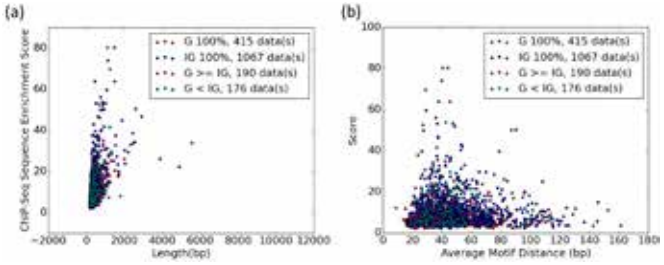

Figure 2. Relationships among different parameters characterizing the binding regions. (A) Relationship between sequence length and ChIP-Seq sequence enrichment score. (B) Relationship between ChIP-Seq sequence enrichment score and average motif distance. G and IG represent genic region and intergenic region, respectively. "G 100%" means that the DNA strand corresponds entirely to genic regions. "IG 100%" means that the DNA strand corresponds entirely to intergenic areas. "G ≥ IG" and "G < IG" mean that the genic regions overlapping with the DNA strands are, respectively, longer or shorter than the intergenic regions.